Achievement of Market-Friendly Initiatives and Results Program
(AMIR 2.0 Program)

Funded by U.S. Agency for International Development

Quantitative Analysis

Introduction to Basic Concepts and Comments on Learning Outcomes
(CFA II, 2003)

By
Ronald E.Copley, Ph.D., CFA
President
Copley Investment Management

Contract No. 278-C-00-02-00210-00

2003

Quantitative Analysis

Introduction to Basic Concepts and Comments on Learning Outcomes (CFA II, 2003)

Note: Candidates will need to refer to the Learning Outcomes in your Study Guide for wording of each Learning Outcome (LO). I have provided you with comments on each LO in each study session, but have not duplicated the LOs in these notes.

by

Ronald E. Copley, Ph.D., C.F.A.
President
Copley Investment Management
5218 Wrightsville Ave.
Wilmington, NC 28403 USA
T: 910-452-7147
F: 910-395-6849
E: RonCopley@AOL.com

Table of Contents

Executive Summary

Linear Regression and Correlation

Correlation analysis is used to portray the relationship (correlation) between 2 or more variables. Originated by Karl Pearson (1900) the coefficient of correlation measures the strength of the relationship between two sets of interval scaled or ratio scaled variables. It is often referred to as Pearson's r and as the Pearson product-moment correlation coefficient. If there is no relationship between the two sets of variables r will be 0. This demonstrates that the relationship is weak. As the coefficients approaches 1 the relationship strengthens. For a coefficient of 1 there is a perfect relationship. Scatter diagrams with divergent points indicate little correlation, while diagrams with data points grouped tightly around the center of the data indicate strong correlation.

The Coefficient of Determination is computed by squaring the correlation coefficient. This is a proportion or percent of the variation in on variable associated with or accounted for by the variation in another variable. A high correlation coefficient or coefficient of determination does not imply causality. The relationship could be spurious.

Regression analysis is a technique used to develop the equation for a straight line. The regression equation is used to estimate y based upon x. It is the mathematical equation used to define the relationship between two variables. The least squares principle determines a regression equation by minimizing the sum of the squares of the vertical distances between the actual y values and the predicted values of y.The assumptions underlying linear regression are:

1. For each value of x, there is a group of y values and these y values are normally distributed.
2. The means of these normal distributions of y values all lie on the straight line of regression.
3. The standard deviations of these normal distributions are equal.
4. The y values are statistically independent. This means that in the selection of a sample, the y values chosen for a particular x value do not depend on the y values for any other x value.

Confidence intervals report the mean value of y for a given x. Prediction intervals report the range of values for a particular value of x.

William Gossett developed the concept of t in the early 1900's. He observed that z was not correct for small samples. This reflects the fact that smaller samples carry greater error.

Multiple Regression and Correlation

The equation for a multiple regression analysis takes the following form:

$$y = a + b_1x_1 + b_2x_2 + b_kx_k$$

The multiple standard error of the estimate is the measure for estimate errors in multiple regression.

The formula for degrees of freedom in multiple regression is:

$$n - (k + 1),$$

where n is the number of observations and k is the number of independent variables. The difference

between the estimated and actual value is called the residual.

Assumptions of multiple regression analysis:

1. The independent and dependent variables have a linear, or straight-line relationship.
2. The dependent variable must be continuous and at least interval-scale.
3. The variation in the difference between the actual and the predicted values must be the same for all fitted values of y. That is, (y - y') must be approximately the same for all values of y'. When this is the case the differences exhibit homoscedasticity. Further, the residuals, computed by y - y', should be normally distributed with a mean of 0.
4. Successive observations of the dependent variable must be uncorrelated. Violation of this assumption is called autocorrelation. Autocorrelation often occurs in time series.

The coefficient of multiple determination R^2 is the percentage of variation explained by the regression.

A global test determines whether the dependent variable can be estimated without relying on the independent variables or could R^2 occur by chance.

h_0: $b_1 = b_2 = ...b_k$
h_a: not all b's are equal to zero

The F distribution is used to test the null hypothesis that all multiple regression coefficients are equal to zero. The decision rule is to accept the null hypothesis if the computed value of F is less than the tabular value. If the computed F is greater than the tabular value reject the null hypothesis.

Use the t test to determine if individual regression coefficients are equal to zero. Accept the null hypothesis (b = 0) if the computed t is less than the tabular t.

Qualitative variables may be used in a multiple regression equation. They are called dummy variables.

Stepwise regression enters independent variables into the regression equation in the order in which they explain the most variation (increase R^2) in the independent variable.

Residuals should be normally distributed. Histograms or plots may be used to show that there are no trends or patterns.

<u>An Introduction to Decision Making Under Uncertainty</u>

There are three components to any decision-making situation (1) Choices available, (2) States of Nature, and (3) the payoffs. The expected payoff is equal to the probability of a state of nature occurring multiplied by the payoff.

All combinations of decision alternatives and states of nature result in a payoff table.

There are several criterions for making the optimal decision.

a) The Expected Monetary Value (EMV) approach requires computing the expected value for each decision and selecting the largest EMV.
b) An opportunity loss table can be constructed by taking the difference between the optimal decision for each state of nature and the other decision alternative. The difference between the optimal decision and any other decision is the opportunity loss or regret. The Expected

Opportunity Loss (EOL) is combined with the probabilities of the various states of nature for each decision alternative to determine the expected opportunity loss.

The strategy of maximizing the minimum gain is referred to as the maximum. The strategy of maximizing the maximum gain is called maximax. The strategy that minimizes the maximum regret is called minimax.

The expected value of perfect information (EVPI) is the difference between the expected payoff if the state of nature is known and the optimal decision under conditions of uncertainty.

Sensitivity analysis evaluates the effect of various probabilities for states of nature on the expected values.

Decision trees are useful for structuring various alternatives. They depict the courses of action and the possible states of nature.

Opportunity loss is the payoff that may be lost because the exact state of nature was not known.

Fundamental Factor Models

Harry Markowitz developed the idea that risk could be quantified using the variance or standard deviation of a stocks return. This concept is referred to as total risk. The total risk of a portfolio is not the sum of total risk of each stock in the portfolio but rather the covariance or correlation between each pair of stocks in the portfolio. Markowitz defined an efficient portfolio as one which minimized the total risk of a portfolio for any level of expected return. Subsequent analysis of risk resulted in the development of two principles:

1. Systematic risk is the only risk investors are rewarded for assuming in efficient capital markets. Systematic risk cannot be diversified away by investors.
2. Unsystematic risk may be diversified away and earns no return in efficient markets.

Asset pricing models show the relationship between the expected return for a stock and risk factors. The two dominant asset pricing models are: the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Model (APM). The CAPM holds that there is only one form of systematic risk. This risk is movement of the market in general. In theory the market in general is defined as all assets. Practically it is the rate of return on a broad-based market index such as the S&P 500. CAPM holds that the expected return on an individual asset is a positive linear function of its index of systematic risk as measured by Beta. Only an assets beta determines its expected return.

There are two approaches to estimating Beta. First, using monthly or weekly returns and running a linear regression against some market index may estimate it. The Beta statistic generated using this approach is referred to as the historical beta. The regression equation is called the market model. The problem with the historical beta is that it is based upon historical data while the fundamental attributes of a company will change in the future. The second approach to estimating beta attempts to resolve this problem. The analyst will first calculate the historical beta for all firms using the market model. Next the fundamental factors, which affect the betas for firms, will be determined. Finally, given the estimates for the parameters, a company's beta is estimated using the fundamental attributes. The resulting beta is called a fundamental beta. The weakness of this approach is that it is assumed all firms are equally impacted by a fundamental factor.

Stephen Ross developed the Arbitrage Pricing Model in 1976. The APM holds that there may be more than one systematic risk. APM does not specify the fundamental factors. The APM states that investors want to be compensated for all factors that systematically affect the return of a stock. The compensation for assuming risk is the sum of the products of each factor's systematic risk. If the only factor in the APM is market risk the APM reduces to the CAPM. CAPM and APM are equilibrium models that tell us the relationship between risk and expected return. Factor models are empirically derived models that seek to identify the risk factors that explain stock returns.

There are three types of factor models: statistical factor models use historical and cross-sectional data on stock returns to derive factors which best "explain" stock returns, macroeconomic factor models use historical stock returns and macroeconomic data to determine the economic variables that best "explain" stock returns, fundamental factor models use company in industry attributes as well as market data to "explain" stock returns. Fundamental Factor Models (FFM) use company and industry attributes and market data as basic inputs to the model. Important factors are called raw descriptors. The output of a FFM is the expected return for a stock after adjusting for all of the risk factors. This return is called the expected excess return. From the expected excess return for each stock, a weighted average expected excess return for a portfolio comprised of a number of stocks can be computed. Similarly the sensitivity of a portfolio to a given risk factor is a weighted average of the factor sensitivity of the stocks in the portfolio. The set of factor sensitivities is the portfolio's risk exposure profile. Portfolio Managers can compute expected excess returns and develop risk exposure profiles for a market index in order to measure performance.

Study Session 3, Investment Tools: Quantitative Methods for Valuation, DeFusco, McLeavey, Pinto, and Runkle, Quantitative Methods for Investment Analysis.

1A) Hypothesis Testing, Ch. 7
Selected end-of-chapter problems: 1-5.

Steps in hypothesis testing include: (1) state hypothesis, (2) identify test statistic and its probability distribution, (3) specify level of significance, (4) state decision rule, (5) collect data and perform calculations, (6) make statistical decision, (7) make economic decision pg 317).

Null hypothesis is the hypothesis to be tested. It is stated as equality. Alternative is hypothesis accepted when null is rejected. It is stated as complement to null (pg 318).

Chose null to be what is considered true. Choice of alternative is complement of null.

A test statistic is a quantity calculated on the basis of a sample, whose value is the basis for deciding whether to reject or not reject eh null hypothesis.

Significance level is the probability of a Type I error (rejecting the null when it is true) in testing a hypothesis.

Power of the test is the probability of correctly rejecting the null (pg 321).

Decision rule state when to accept or reject the null hypothesis. When rejecting the null, the result is statistically significant (pg 322).

Confidence intervals represent an alternative method of tests of significance. When the hypothesized value of the population parameter under the null is outside the corresponding confidence interval, the null hypothesis is rejected (pg 324).

z-test is appropriate for mean testing when the sample size is large since the Central Limit Theorem says that the distribution of the mean is normal when n is large.

The test statistic for equality of two population means from independent samples is a t-statistic. (pg 334).

Test statistic for a paired comparison (samples are not independent) test is a t-statistic (pg 338).

Null hypothesis is that the hypothesized value equals some specific value; the alternative hypothesis is that they are unequal. If the test statistic (chi-square) falls outside the acceptance ranges, the null hypothesis is rejected, and vice versa (pg 341).

The null hypothesis is that the two variances are equal. The alternative is that they are not equal. Reject the null if the test statistic falls outside the acceptance region, and vice versa (pg 344).

Learning Outcomes
(Please refer to your Study Guide for wording of LOs. The following are comments I have made for each LO).

a) For a one-tail-test, the alternative hypothesis gives rejection in only one tail of distribution. For a two-tail test, the alternative gives rejection in two tails of the distribution.

b) Type I error is rejecting the null when it is true and a Type II error is not rejecting null when it is false (Table 7-1, pg 321).

c) The probability of a Type I error is the level of significance. For example, a level of significance of .05 for a test means that there is a 5 percent probability of rejecting a true null hypothesis. By decreasing the probability of a Type I error, we increase the probability of a Type II error. The only way to reduce the probability of both types of errors simultaneously is to increase the sample size, n (pg 321).

d) Statistical decision is based on empirical data contained in the test. An economic decision is based on the statistical decision, but also all economic issues pertinent to the decision (pg 325).

e) p-value is the smallest level of significance at which the null hypothesis can be rejected (325).

f) Test statistic is t when population variance is unknown (pg 327), but z when population variance is known (pg 330).

g) The null hypothesis tests whether the population mean is equal to, less than or equal to, greater than or equal to a specified level. The alternative hypothesis is the complement of the null. Reject the null if the test statistic falls outside the acceptance region (pg 330).

h) Null hypothesis is Ho: mean1 – mean2 = 0. Rejection of the null is when the test statistic falls outside the acceptance region (pg 335).

i) Null hypothesis mean1difference = mean2difference where mean1difference equals difference between means of paired samples (Table 7-5. pg 338).

j) Choice depends on whether sample is independent or not. Equality of two population means assumes independent samples whereas paired comparison assumes samples are not independent.

k) Test statistic is a chi-square (pg 341).

l) The test statistic is an F-test, which is the ratio of sample variances (pg 343).

m) Parametric tests assume that the test statistic used in the hypothesis is drawn from a population that has a certain distribution, such as normality (pg 345).

1B) Correlation and Regression, Ch. 8
Selected end-of-chapter problems: 5, 11, 17.

Scatter plot is a graph that shows the relationship between the observations for two data series in two dimensions (pg363).

If the correlation coefficient equals 1.0 (-1.0), the relationship between two variables is perfectly positive (negative). That is, by knowing the independent variable, you will know exactly the dependent variable.

Dependent variable (left hand side of equality sign) derives its value from the independent variable (right hand side of equality sign). The independent variable explains changes in the dependent variable (pg 379).

In a two-variable regression, the slope measures the degree of the relationship. A slope greater than 1.0 means a small change in the independent variable results in a larger change in the dependent variable, and vice versa. The intercept gives the value of the dependent variable when the independent variable equals zero (pg 380).

Assumptions of linear regression are: (1) linear relationship, (2) independent variable is not random—this is a simplifying assumption, but not essential, (3) expected value of error term is zero, (4) variance of error term is constant across all observations of independent variable, (5) error terms are uncorrelated, (6) error term is normally distributed. All of these assumptions are made for convenience, but we have sophisticated techniques for adjusting the regression technique (pg 383).

Confidence intervals are intervals that we believe include the true parameter value being tested and depend on the SEE (pg 391).

Learning Outcomes

a) Covariance is a measure of linear relationship between two variables. It has no boundaries (pg

b) Correlation coefficient is a measure of linear relationship between two variables. It is bound by –1 and +1 (pg 365).

c) The null hypothesis is that r = 0. The alternative is r does not equal 0. If the test statistic, t, falls outside the acceptance region, reject the null hypothesis, and vice versa (pg 377).

d) Outliers are small numbers of observation at either extreme of a sample. Outliers weaken the correlation coefficient (pg 368).

e) Spurious correlation is not based on any theoretical relationship (pg 370).

f) Standard error of the estimate measures the uncertainty of the regression equation when estimating the dependent variable. The larger the error term, the larger the SSE, and vice versa (pg 386). Coefficient of determination measures the fraction of the total variance in the dependent variable that is explained by the independent variable (pg 388).

g) The test statistic is a t since the regression equation is estimating more than one variable. In the case of a simple regression, the t-statistic has n-2 df since we are estimating a slope and an intercept. For each additional variable being estimated, we lose a df (391).

h) The null hypothesis says the slope, for example, equals zero: B=0. The alternative says not equal. Rejection of the null occurs when the test statistic falls outside the acceptance region (pg 391).

i) Regression coefficients allow estimation of the dependent variable given values of the independent variable(s) (pg 380).

j) Predicted values come from the regression equation that depends on the accuracy of the regression coefficients (pg 399).

k) The confidence interval for the estimated (or predicted) value of the dependent variable depends on the standard deviation of the forecast error, which depends on the SEE, number of observations, value of independent variable, estimated mean of the independent variable, and the variance of the independent variable (pg 400-401).

l) ANOVA is a statistical procedure for analyzing the total variability of a set of data into components that can be attributed to different sources. It is useful in determining whether the independent variables can explain variation in the dependent variable (pg 396).

m) In a simple regression, the F-statistic duplicates information contained in the t-statistic. In multiple regression, the two are not duplicative. The F-statistic tests whether all the slope coefficients in the regression equation equal zero (pg 396).

n) Limitations of regression analysis include: (1) relations can change over time, (2) widespread knowledge allows other analysts to take advantage of analysis, (3) many assumptions stated in j above are violated in practice (pg 403).

1C) Multiple Regression and Issues in Regression Analysis, Ch. 9
Selected end-of-chapter problems: 1, 7, 15.

Multiple regression involves two or more independent variables to explain changes in the dependent variable (pg 429)

Statistical significance for each independent variable is determined by use of the t-statistic since we do not know the true value of the population variance. Do not confuse this test with the F-test that determines the significance of the entire equation (pg 431).

The two types of uncertainty are (1) uncertainty in the regression model itself, as reflected in the standard error of the estimate, and (2) uncertainty about the estimates of the regression model's parameters (pg 436).

Learning Outcomes

a) The relationship between a dependent variable and multiple independent variables is linear. Changes in the independent variables are used to explain changes in the dependent variable (pg 429).

b) The two-tail null hypothesis is that the coefficient (intercept and slopes) = 0. The alternative is that they are not equal to zero. Calculation of the t-statistic: (estimated value of coefficient – hypothesized value of coefficient) / standard deviation of coefficient. If the t value falls outside the acceptance region for a two-tail test, reject the null. If the null is less than or equal to zero (one-tail test), the acceptance region is to the left of the critical value, and vice versa (pg 431).

c) The confidence intervals for multiple regressions are the same as for simple regressions (pg 391).

d) Assumptions are: (1) linear relationships, (2) independent variables are not random, (3) expected value of error term is zero, (4) variance of error term is constant across all values of the independent variables, (5) error terms are uncorrelated, (6) error term is normally distributed (pg 432).

e) The standard error of the estimate depends on the sum of the squared residuals (pg 435).

f) The standard error of the estimate equals the sum of the squared residuals / (# of observations – 1 + # of independent variables) (pg 435).

g) Given the regression equation, the predicted value of a dependent variable is determined by multiplying each independent variable by its regression coefficient and summing (pg 436).

h) The F-statistic is used to determine significance of the overall model. The null hypothesis is the all the slope coefficients simultaneously equal zero (pg 437).

i) R2 gives the percentage of variation explained by the regression equation. R2 increases as additional variables are added to the model. Adjusted R2 does not automatically increase when another variable is added to the regression; it is adjusted for df (pg 439).

j) An ANOVA table provides information on the explanatory power of a regression and the inputs for an F-test of the null hypotheses that the regression coefficients all equal zero (pg 437).

k) Dummy variables (independent) take on a value of 1 if a particular condition is true and 0 if that condition is false (pg 439).

l) Heteroskedasticity violates one of the assumptions that the variance of the error term is constant across all observations of the independent variables. In other words, the variance is not constant. Unconditional heteroskedasticity occurs when the heteroskedasticity is not correlated with the independent variables in the model. This is not a problem for statistical inference. Conditional heteroskedasticity is when heteroskedasticity is correlated with the independent variables.

m) Serial correlation occurs when regression errors are correlated across observations. Note that serial correlation can occur in either time series or cross-sectional data. The effect is to misestimate the

standard error of the estimates that, in turn, impact hypothesis testing (pg 450).

n) Heteroskedasticity can be observed via a scatter diagram (graph) and corrected in one of two ways: (1) generalized least squares, or (2) computing robust standard errors. Neither method is explained in the text. Serial correlation is tested by use of the Durbin-Watson statistic. It can be corrected in one of two ways: (1) adjust the coefficient standard errors, and (2) modify the regression equation to eliminate the serial correlation (pg 453).

o) The DW statistic approximately equals 2(1-r) for large sample sizes. If DW equals 2.0, error is not serially correlated (pg 452).

p) Multicollinearity occurs when two or more independent variable are highly (but not perfectly) correlated with each other, which causes a problem for interpreting regression output (pg 457).

q) Qualitative dependent variables are dummy variables used as dependent variables instead of as independent variables. This technique is used when the outcome is either success or failure (pg 460).

r) The economic meaning of a regression output must have a theoretical basis and not be spurious.